

AD-A162 458

GEONAMES PROCESSING SYSTEM FUNCTIONAL DESIGN
SPECIFICATION VOLUME 1 AUTOM. (U) NAVAL OCEAN RESEARCH
AND DEVELOPMENT ACTIVITY NSTL STATION NS..
G LANGRAN ET AL. MAR 85 NORDA-98

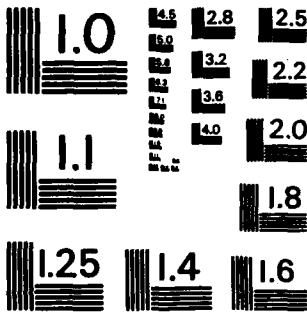
1/1

UNCLASSIFIED

F/G 9/2

NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD-A162 458

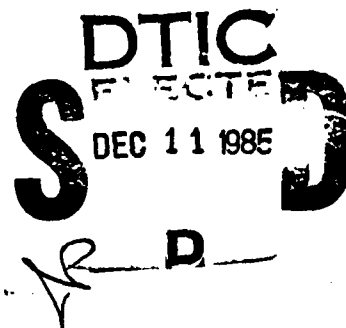
Naval Ocean Research and Development Activity
March 1985

Report 98



Geonames Processing System Functional Design Specification

Volume 1: Automated Alphanumeric Data Entry System

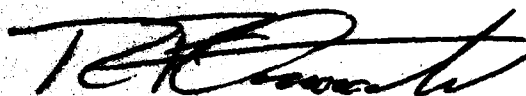


DTIC FILE COPY

Gail Langran
Mapping, Charting, and Geodesy Division
Ocean Science Directorate

Anne Downs
Barry Glick
Warren Schmidt
PAR Technology Corporation
McLean, Virginia

...and the digital procedures to store, retrieve, and edit geographic names
...DBA's stated goal is a 50-100 million name digital
...edit and format data, and prepare names overlays
...system design study late in FY82. This report is
...the functional design of the digital geographic names



R. P. Onorati, Captain, USN
Commanding Officer, NORDA

This NORDA Report was prepared to meet style requirements of the sponsor.

EXECUTIVE SUMMARY

In FY82, the Pattern Analysis Branch, Mapping, Charting and Geodesy Division of the Naval Ocean Research and Development Activity (NORDA) began a subtask for the Defense Mapping Agency (DMA) entitled, "Advanced Type Placement and Geonames Database System Development." This effort will develop systems to address four interrelated aspects of computer-assisted geographic names processing as follows.

- **Data Capture:** digital capture of names and named feature information from analog sources such as maps, gazetteers, and other data sources.
- **Data Management:** development or adaptation of a Data Base Management System for a very large product-independent set of world geographic names and their descriptors. This data base will support a variety of DMA products including maps, charts, and gazetteers.
- **Data Manipulation and Editing:** in support of toponymic research, advanced word processing for text containing diacritics and special symbols, document formatting, data file searching, and statistics generation.
- **Product Generation:** digital text placement on maps, gazetteers, and other DMA products with the associated data selection, formatting, scaling, and type assignment.

This Geonames Processing System subtask is scheduled for performance during FY82-FY89. During the first year (FY82) an initial Comprehensive Coordination Plan (CCP) was generated for the technical description of the above automated names capability (NORDA Technical Note 189). The second stage of planning built on the CCP and DMA responses to the CCP to generate the present five volume set of Functional Design Specifications, one for each subsystem and one to describe requirements that are mutual to all four subsystems.

This report is a functional analysis of the task of mass digitization of geonames and their associated attributes for storage in the data base. An analysis of task elements is followed by a discussion of the current state of technology and development options. Technology has not yet reached a level where it can adequately respond to the data capture requirements of the Geonames Processing System. The information presented in this report will provide a basis for deciding the course of future development.

ACKNOWLEDGMENTS

This work was sponsored by DMA under Program Element 64701B, subtask title, "Geonames Processing System." Mr. Dennis Franklin and Lt. Col. Tom Baybrook, both of DMAHQ/STT, shared project management duties during the writing of this report. Their help in communicating with DMA's production centers and providing information on DMA production methods was instrumental to this functional design. Dr. Don Durham, head of NORDA's Mapping, Charting, and Geodesy (MC&G) Division, and Dr. Charles Walker, head of the MC&G Division's Pattern Analysis Branch, contributed valuable advice and assistance.

TABLE OF CONTENTS

INTRODUCTION

v

1.0	AUTOMATED ALPHANUMERIC DATA ENTRY SYSTEM OVERVIEW	1-1
2.0	AADES ALTERNATIVES AND RECOMMENDED CONFIGURATION	2-1
2.1	Organizational Strategies	2-1
2.2	Technical Strategies	2-2
2.3	Recommended AADES Configuration	2-3
3.0	HARDWARE REQUIREMENTS	3-1
3.1	Map Scanner	3-1
3.2	Graphics Workstation	3-1
4.0	SYSTEM FUNCTIONS AND SOFTWARE	4-1
4.1	Preprocessing	4-1
4.2	Raster Scanning	4-1
4.3	Character Isolation and Recognition	4-2
4.4	Word Construction	4-2
4.5	Duplication Review	4-3
4.6	Generic Elimination	4-3
4.7	Feature Association	4-3
4.8	Position Capture	4-3
4.9	Attribute Capture	4-4
4.10	Quality Control	4-4
4.11	Output Processing	4-4
4.12	File Management	4-5
4.13	Job Management	4-5
5.0	DATA SETS	5-1
5.1	Lookup Tables	5-1
5.2	Workfiles	5-1
5.3	Data Base Update File	5-1

APPENDIX A: PREVIOUS WORK IN NAMES DATA CAPTURE

A-1

APPENDIX B: REFERENCES

B-1

Accession For	
NTIS CRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution	
Availability Codes	
Dist	Availability for Special
A-1	



FIGURES

<u>Figure</u>		<u>Page</u>
5-1	Standard Data Transfer Record	5-2
A-1	USGS Data Capture Rate	A-2

TABLES

<u>Table</u>		<u>Page</u>
1-1	Data Entry Timing	1-2
2-1	Average Map Timing	2-2
A-1	USGS Data Capture Rate	A-2

INTRODUCTION

a. Organizations

Defense Mapping Agency Headquarters (DMAHQ)
U.S. Naval Observatory
Washington, D.C.

Defense Mapping Agency Hydrographic/Topographic Center (DMAHTC)
6500 Brookes Lane
Washington, D.C.

Defense Mapping Agency Aerospace Center (DMAAC)
3200 South Second St.
St. Louis, Missouri

b. Scope

The purpose of this Functional Design Specification is to state the system requirements to be satisfied which will serve as a basis for mutual understanding between the user and the developer.

The Geonames Processing Subsystems discussed in this document will be referred to by their acronyms: ASP (Advanced Symbol Processing); ATP (Advanced Type Placement); GNDB (Geographic Names Data Base); and AADES (Advanced Alphanumeric Data Entry System).

c. Background

In FY82 the Pattern Analysis Branch, Mapping Charting, and Geodesy Division of the Naval Ocean Research and Development Activity (NORDA) began a subtask for the Defense Mapping Agency (DMA) entitled "Advanced Type Placement and Geonames Data Base System Development," a project encompassing the digital capture, storage, edit, and display of geographic names. The subtask in its current form is an amalgamation of four previous DMA requirements for the independent development of: a geographic names data base; a system design for geographic names data capture; advanced symbolic processing; and automated type placement for maps (see Appendix B for the original Requirements Statements). A Comprehensive Coordination Plan was submitted by NORDA as a preliminary definition of the overall Geonames Processing System subtasks and their interfaces.

d. Description

The complete Geonames Processing System is comprised of four components, as follows (see Figure i-1).

1. Automated Alphanumeric Data Entry System (AADES). This subsystem provides a means of high-volume names data capture. World geonames with corresponding locations and attributes will be captured from both tabular and map/chart sources using raster scan and optical character reading technologies. The AADES subsystem will convert alphanumeric data into computer readable form

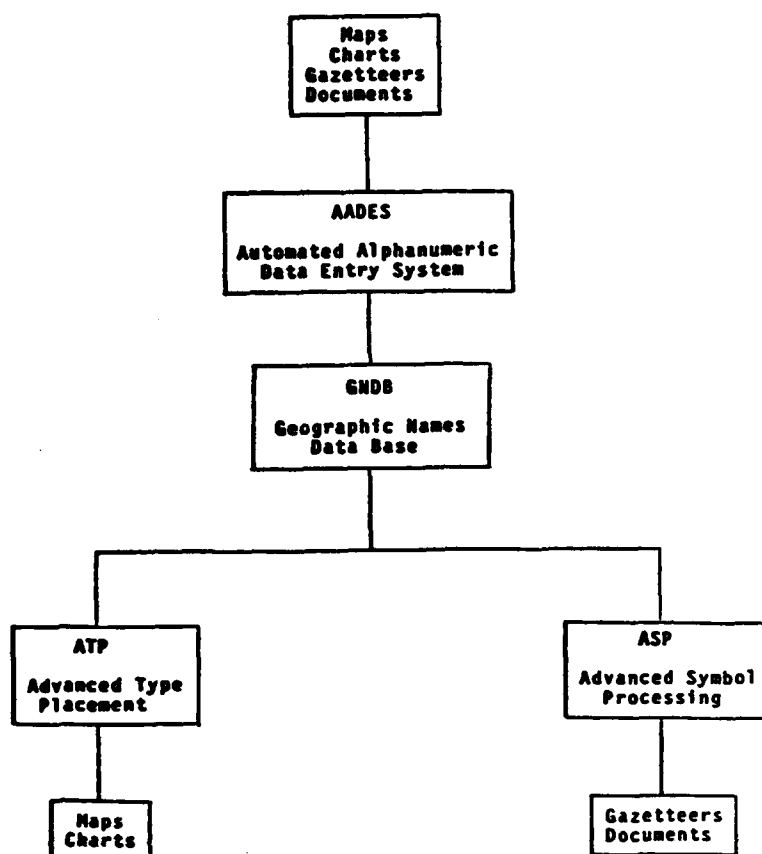


Figure i-1. Geonames Processing System overview

with a 99% accuracy rate. It will require minimum operator intervention, provide automated error checking, and result in clean data files for supervised merging with the GNDB.

2. Geographic Names Data Base (GNDB). The GNDB will respond to the storage requirements for world geonames and their descriptors in non-product oriented file formats. It will provide extensive query capabilities to support data base updates, chart and gazetteer compilation, and toponymic research. The ultimate size of the GNDB will be 50-100 million geonames.

3. Advanced Symbol Processing (ASP). International geonames comprised of diacritics and special symbols require specialized hardware and software in order to carry out normal data access and manipulation as well as text editing operations. The ASP subsystem will provide alphanumeric edit and display of world geonames in addition to advanced word processing capabilities such as sorting, searching, and formatting.

4. Advanced Type Placement (ATP). This subsystem will automate the production of map names overlay by exploiting electronic display technology and the rule-based nature of cartographic names placement. This subsystem includes automated utilities for names selection, type composition, type placement, virtual map display, and interactive graphic edit.

The Geonames Processing System described above will respond to a major need: it will integrate the splintered DMA names processing system into the digital map production pipeline and coordinate all geonames processing activities. The obvious benefits are increased production rates and lower costs in geonames processing. It is also expected that overall accuracy and coverage will improve since improved techniques and helpful utilities will increase the level of productivity of toponymic researchers. A less apparent benefit is the ease with which future innovations and improvements can be made to a fully digital system.

This Functional Design Specification addresses the development of the Automated Alphanumeric Data Entry System (AADES). The AADES subsystem is the least defined of the four subsystems due to the relative inability of today's technology to deal with mass entry of 50 million geonames from largely chart source. Thus a functional analysis is submitted in order to provide a basis for determining the direction to be taken with AADES development.

e. Applicable Documents

The following references provide a summary of the basis for the Geonames Processing Subtask development.

1. "Advanced Type Placement and Geonames Database: Comprehensive Coordination Plan." NORDA Technical Note 189, January 1983.
2. "A Prototype Geographic Names Input Station for the Defense Mapping Agency." Paper presented at Auto Carto IV by D.R. Caldwell and D.E. Strife, September 1982.
3. "Names Type File System." Consulting Report for the U.S.A.E.T.L Project #POO13. April 1983.
4. "Development of an Automated Cartographic Capability." The Final Report of the Automated Cartography Task Force, Defense Mapping Agency Hydrographic/Topographic Center, April 1982.
5. "The Feasibility of Establishing an Automated Chart Production Process." The Defense Mapping Agency Hydrographic/Topographic Center, October 1982.

f. Limitations

The individual subsystem descriptions are functional and not physical definitions. i.e.:

1. a given function required by a given subsystem may not be performed upon the hardware logically associated with the subsystem,
2. one software module may serve several of the subsystem functional requirements.

Thus there is some redundancy in the functions and data sets specified in this five volume set of design specifications. This is intentional. It is felt that complete functional analysis of each separate subsystem will allow full system definition. Physical (hardware and software) synthesis will be accomplished and described by the Implementation Plan at a later date.

1.0 AUTOMATED ALPHANUMERIC DATA ENTRY SYSTEM OVERVIEW

The Defense Mapping Agency (DMA) houses an analog data base of worldwide geographic names. This data base resides on cards, in gazetteers, and on maps and charts. A small part of the data base is duplicated on magnetic tape. To promote names processing efficiency, this analog data base will be converted to a digital data base supported by a state-of-the-art data base management system. The eventual goal is a digital data base containing 50-100 million geonames and their attributes.

Digitized names data will be used for toponymic research report and gazetteer production, and map and chart annotation. Toponymic research requires all possible feature attributes and all known variant spellings and aliases. Map annotation requires accurate feature coordinates for placement, and attributes to decide which features to name and what type style to use.

Published maps and charts are the richest source of names data and are the only source of accurate named feature coordinates. The current state of the art, unfortunately, provides only rudimentary means for capturing names data from graphic sources.

The Automated Alphanumeric Data Entry System (AADES) described in this report will provide DMA with a means of high-volume geonames data entry from map sources. Source maps will include DMA map series and non-DMA maps (foreign and domestic). Captured information will be stored in a sequential file to be merged with the Geographic Names Data Base (GNDB). This file will include information on the geoname (specific and generic components, if applicable), the position of the feature to which the name refers, the feature's designator or class, the source document, the country (or countries) the named feature is in, the feature's attributes, and the non-Romanized name (if applicable). *Keywords: maps, requirements*

Keyboard entry and manual digitization are available technologies for a names data capture system. However, these technologies cannot be upgraded for faster processing. Voice entry is another available and applicable technology. However, DMA has requested that voice entry not be used. Optical character recognition (OCR) offers the most rewards if implemented. OCR technology, however, has not been commercially implemented for text with the diacritics and special symbols found in geonames.

This report focuses on rapid and cost-effective names digitization from graphic sources. Speed is imperative because of the data base's size requirement. Table 1-1 shows the amount of time it would take to capture a 50-million-name data base given the time required to capture a single name and its attributes. Data base maintenance may compete for the same resources and slow data capture even further. It is evident that even small timing improvements will yield cost savings for years to come.

The system must be interactive if it is to be operable in this decade. An interactive AADES can be implemented using currently available hardware and software. Yet to maintain a work flow of this magnitude, the final version of AADES must be highly optimized and thoroughly tested. Considerable time and cost savings may be achieved by automating certain AADES functions. Character recognition is a major component of automating the process of reading names from maps and charts. Yet background noise, irregular letter spacing and orientation, nonstandard type styles, diacritics, and hand-printed characters make recognition more difficult. Also difficult to automate are connecting letters into words, distinguishing generic map names (such as "rocks," "sand") from placenames, and relating a name to its named feature. A balance must be struck between man and machine for the most accurate and cost-effective AADES strategy. The next section recommends an AADES configuration after reviewing the alternatives. Hardware, software, and data sets for the recommended system are discussed in sections 3, 4, and 5, respectively. Appendix A reviews operational geonames capture programs and systems. Appendix B discusses relevant hardware and software technologies.

Because optimal systems cannot be procured off the shelf today, AADES must be implemented as a prototype.

Table 1-1. Data Entry Timing.

Names to goal
50,000,000

Days/Year 365	Wkends & Holidays 114	Down Time 5%
Working Days 238	Hrs/Day 7	Cost/Manyear \$75,000
Mins/Name 3.0000 1.0000 0.5000 0.2500 0.0167	Man/ys to Goal 1497.77 499.26 249.63 124.81 8.32	Human Cost \$112,332,624 \$37,444,208 \$18,722,104 \$9,361,052 \$624,070

* Minutes per name is the average number of minutes spent interactively identifying a name and its attributes. Batch portions of the task are not included.

2.0 AADES ALTERNATIVES AND RECOMMENDED CONFIGURATION

The selection of a names data capture approach must be based on organizational and cost considerations. The organizational strategy will dictate how usable and useful DMA's names data base will be in the near future (25 years). Cost considerations are a tradeoff between hardware and personnel expense. This section discusses these issues and recommends a technical approach to AADES.

2.1 Organizational Strategies

2.1.1 "Build as You Go"

In the course of production, names data can be digitized from maps and charts and entered into the GNDB. This practical data capture method would require approximately 26 man/hours per map or chart (based on an average of 500 names/chart if each name takes three minutes to capture (Table 2-1). Advantages of this method are that it minimizes system-specific personnel and hardware requirements, and the slow pace of data capture would promote orderly data base loading.

Unfortunately, capturing names data in the course of production would add to production time and result in poor data base coverage. The GNDB would neither contribute to production efficiency nor would it be a reliable source of geonames for at least 10 years. Names that do not appear on recently updated maps will be excluded from the data base until their respective maps are scheduled for revision. These problems could reduce what was intended to be a powerful reference tool to a mere archive, and could lead to disenchantment with the entire geonames processing system.

2.1.2 Dedicated Names Data Capture

Each government agency that has built a names data base has opted for a bulk, one-time data capture effort (see Appendix A). If names are to be captured in an independent effort, they may be captured by a small team of in-house personnel or by contractors. Loading the captured data into the data base must be performed by DMA's toponymic staff. In either case, the effort should be physically located at DMAHTC to take advantage of AADES hardware and software in all processing stages.

Names data requirements must be prioritized for a dedicated capture effort. One alternative is to prioritize data capture as follows.

- Level one coverage. Names in areas with the highest priority are captured first at 1:50,000 scale, 100% accuracy.
- Level two coverage. Names in areas with second-order priority are captured next at 1:250,000 scale, with slightly lower accuracy levels tolerated.
- Level three coverage. Names in areas with third-order priority would be captured last from 1:500,000 scale source materials. The source scale alone would lower accuracy levels, but would provide general placename coverage for digital applications.

2.1.3 Combined Effort

Combining production-oriented names digitization with a dedicated capture effort is the recommended strategy. Because of the specialized equipment and the long-term nature of the effort, it is also recommended that DMA personnel be used. If data capture is contracted, it should be by area at a fee per correct name. A reasonable fee can be established once the speed of capture using AADES is known. Digitized data sets can be sampled to determine the approximate number of correct names.

Table 2-1. Average Map Timing.

Names on map 500		
Days/Year 365	Wkends & Holidays 114	Down Time 5%
Working Days 238	Hrs/Day 7	Cost/Manyear \$75,000
Mins/Name 4.0	Man/hrs for map 35.09	Human Cost \$1,577
3.5	30.70	\$1,380
3.0	26.32	\$1,182
2.0	17.54	\$788

2.2 Technical Strategies

The technical strategy dictates hardware/personnel tradeoffs. Tradeoffs become clear when technical alternatives are grouped in terms of their source media—hardcopy (analog) or softcopy (scan digitized).

Softcopy-source and hardcopy-source systems are distinguished by the media from which the name is captured in ASCII form. Keystroked entry, voice entry, and selective scanners (e.g., scanning wands and scanning cursors) capture the name directly from hardcopy source. Alternatively, the entire map can be scan-digitized and the names captured from this softcopy source. The relative merits of each approach are discussed next.

2.2.1 Hardcopy Source

Using hardcopy source materials reduces hardware expense, since there is no need to procure scanners or graphic workstations, both expensive equipment items. However, hardcopy approaches are relatively more labor-intensive and of limited upgradability.

The simplest hardcopy approach is to fasten the source map to a digitizing table, digitize the feature location, and keystroke names and attributes using a keyboard. Preliminary preparation of source materials (see the description of USGS names data capture) speeds processing somewhat. The hardware to perform this task is inexpensive and available.

Some alternative technologies apply to hardcopy source materials. Voice entry of names and feature attributes could slightly improve timing, but DMA has ruled out voice entry. OCR has been implemented on a scanning wand and could conceivably be implemented on a scanning cursor. By scanning individual names from hardcopy maps, only feature attributes would need to be keystroked. However, OCR for maps and for text with diacritics is not straightforward (see later discussions of OCR), and selective scanning introduces problems with wand pressure, tilt, and movement. Before pursuing this alternative, current off-the-shelf capabilities must be demonstrated to and evaluated by project personnel.

2.2.1 Softcopy Source

The softcopy AADES approach scan-digitizes an entire map (unseparated) or a names overlay. After scanning, names are isolated and captured from the raster output using a combination of interactive and automated methods.

Even a fully interactive, softcopy-based AADES has merit. Technology is likely to improve rapidly in this area, since several vendors are actively developing automated blueprint readers. In addition, neck and back discomfort could prevent analysts from operating a digitizing table for long periods, but interactive workstations have been engineered with human comfort in mind. A dedicated names capture effort could support several such workstations. A "build as you go" effort could employ the same workstations used for other interactive tasks.

Software upgrades to a softcopy-based system can be designed around a number of type characteristics and technologies. Type color, size, and shape help to isolate type in the raster image. Heuristic searches connect letters into words. OCR could convert a large percentage of the character images into ASCII codes. Software dictionaries would help eliminate generic words. Man-machine interaction occurs between steps or throughout processing, the machine requesting approval or correction of automated deductions to date, and permission to proceed to the next processing step.

Because of human factors and upgradability, a softcopy-based AADES is recommended. The proposed strategy is to implement low-risk software that will automate where possible, and add interactive utilities for human intervention where machine efforts fail. For example, software developed in DMA's Auto Carto Feature Identification project can automatically identify standard DMA map and chart symbols from their scanned images. Thus, when standard DMA maps are source materials, AADES could automatically recognize those feature's types and digitize their locations—an estimated 50% of the named features. Unrecognized features would be digitized by the analyst.

2.3 Recommended AADES Configuration

A softcopy-based AADES requires a raster-scanner with one or more graphics workstations. The system's physical layout must provide the user with adequate space to handle large maps. A console must be easily accessible from the scanner to enter processing commands. The number of graphics workstations depends on management decisions on budget, schedule, and space. There should be one command console and several graphics workstations per scanner.

Both the scanner console and the graphics workstation must provide comfortable working environments. The system must supply a friendly user interface. Software should allow the user to override and interact with the system at any point during processing.

The next section describes the proposed system in detail. Following system description, applicable technologies are evaluated and other names digitization efforts are described.

3.0 HARDWARE REQUIREMENTS

This section lists the hardware requirements for the recommended AADES configuration. Critical AADES hardware is a map scanner and an interactive graphics workstation.

3.1 Map Scanner

The scanner must accept maps of varying sizes and materials. It is anticipated that large format maps will need to be scanned; input materials up to 40 inches x 40 inches are accepted by some scanners but scanners that handle larger sizes are rare.

The scanner must have color-recognition capability. Depending on the type of map, from 4 to more than 20 colors may be used. Although text generally appears in only a limited range of colors (such as black, blue, brown, and purple), other color information on the map must be captured to recognize mapped features.

The scanner must allow scanning resolution to be varied under operator or software control. This capability can be used to help separate characters from symbols and background noise.

An array processor is needed to handle the major computational tasks associated with raster-scanned data processing. In particular, raster-to-vector conversion places a heavy computational burden on a host computer. Some firms have developed proprietary "black boxes" for raster-to-vector conversion, usually built around an array processor. Other vendors have incorporated raster-to-vector software into turnkey systems.

Following is a summary of scanner hardware requirements:

- input document size: at least 36 x 48 inches;
- scanning resolution: adjustable 100-1000 lines per inch;
- absolute accuracy: ± 25 microns;
- color detection: at least 12 colors;
- gray-level detection (optional): 64 gray levels;
- input material: paper maps or film (printed, hand inked, or penciled).

3.2 Graphics Workstation

The AADES workstation must include the following:

- Digitizer table (manual digitization):
 - large format;
 - medium precision/resolution;
 - softcopy echo;
 - cursor with at least 12 programmable function keys.
- Color Graphics CRT:
 - 19-inch diagonal;
 - medium resolution (512x512);
 - 12 colors simultaneously;
 - user interface through touch, trackball, etc.

- Optional peripherals:
 - OCR wand;
 - scanning cursor;
 - voice data entry;
 - video imaging system.

4.0 SYSTEM FUNCTIONS AND SOFTWARE

AADES has the following functions:

- preprocessing.
- raster scanning,
- character isolation and recognition,
- word construction,
- duplication review,
- generic elimination,
- feature association,
- position capture,
- attribute capture,
- quality control,
- output processing,
- file management,
- job management.

This section describes the system functions and states their software requirements and upgrade possibilities.

4.1 Preprocessing

Basic source map parameters are entered into an on-line bookkeeping utility maintained by the system. The parameters are used by the system to determine which lookup tables, dictionaries, and software to access. Required data include

- DMA/non-DMA product.
- If DMA:
 - stock number
 - product series
- If not DMA:
 - publisher
 - title
 - type styles used (if known)
 - symbols used (if known)
 - projection
 - scale
 - coverage (minimum and maximum latitude and longitude)
- language(s),
- date(s) of map compilation,
- date of AADES processing,
- classification.

The input software must adapt when some or all of these parameters are unavailable. In the unlikely event that serious data deficiencies exist (e.g., unknown projection or date), it must still be possible to process the map, with caveats inserted in the file so shortcomings are known to data base personnel.

4.2 Raster Scanning

Source maps are scan digitized and their data stored in work files. Maps that are too large for the scanner are scanned in sections then mosaicked into one file by the system. The system per-

forms the necessary analog-to-digital conversion. Color separation software must exist for when map separations are unavailable.

The analyst can zoom, scroll, and display the map image in segments using the interactive workstation.

4.3 Character Isolation and Recognition

A batch subroutine isolates characters in the raster image using size, shape, and color to distinguish them from other map symbols. The scanner must allow resolution to be adjusted by software or by operator. Adjustable resolution aids in separating characters from symbols and background noise.

Optical character recognition (OCR) software attempts to recognize all isolated characters. The OCR must operate independently of orientation. Part of the development effort should analyze the tradeoff between template-matching and handwritten character reading techniques. Font libraries can be stored by AADES for template-matching character recognition, which is likely to be faster. Template-matching, however, may fail when unknown type styles are processed or if diacritics were added by hand. Character recognition algorithms developed for handwritten characters will be useful for recognizing nonstandard fonts and text with diacritics.

Recognized characters are highlighted on the display for analyst review and edit. Unrecognized characters are entered by the analyst. The character, its type style and size, and its orientation are recorded in the AADES workfile.

4.4 Word Construction

After characters are isolated and recognized, a batch process connects them into words. The majority of map names are positioned horizontally, which makes connection of characters into words relatively straightforward. A horizontal search procedure is used that begins in the upper left corner and proceeds as if reading English-language text. The first letter of words using both upper and lower case is easily found. Identifying the first letter of words using all upper case is less simple. The system assumes that an irregularly large space precedes a new word.

Locating the characters of names that are not horizontally positioned will be difficult when a horizontal search procedure is used. The first character located may not be the first letter of a word. Each new line could yield a character belonging to the same word; thus, the same word is redundantly located. Satisfactory solutions to these problems must be found. Man-in-the-loop solutions may be more efficient and effective than special-case program branching.

Automatically recognizing names with inconsistent character orientation and spacing is also problematic. The system must have analyst-invoked heuristic search software for maps with a large number of nonhorizontal names. This software could also be self-invoked by the system when it finds a single character with no horizontal neighbors. Once again, man-in-the-loop solutions should be used if they prove to be more efficient.

The heuristic search software first determines whether or not a character is the first character of a word (meaning that it is lower case) and the character's orientation. The area surrounding each character must be searched for neighboring characters within a set distance. Search proceeds to the left and the right of the character, within an angle 45 degrees plus or minus the character's horizontal bisector. Characters of radically different orientation and characters with different type styles are assumed to belong to other words. The search for neighbors is exhaustive when a character is lower case.

Following batch procedures for constructing words, the map image with a box drawn around each word is displayed to the analyst for review and edit.

4.5 Duplication Review

All duplicate line and area names are reviewed by the analyst, since one feature may be labeled several times depending on its extent. Duplicate point feature names are screened by software; those located more than a given distance apart are assumed to name different features. Unoriginal placenaming should be anticipated and doubtful cases must be referred to the analyst.

4.6 Generic Elimination

On most standard DMA maps, all generics use a single type style that is not used for placenames. For those maps the system culls generics in batch. On other maps, generics are culled using a generic dictionary (selected according to product type and language) or by interactive means. The interactive generic-culling subroutine displays map portions and deletes the names as indicated by the analyst.

A generic component of a proper name (e.g., "river" in Moon River) are considered part of the name itself. If the generic component is identified, software could use it to identify feature type and to search for the named feature.

4.7 Feature Association

Each placename must be matched to its named feature. This process is called "feature association." The initial system should provide software for interactive feature association. Several automated feature association upgrades are feasible, however.

The first step of automated feature association is to search the area surrounding each name for a feature symbol. Occasionally, leader lines can be followed to a feature. Also, reversing the rules of cartographic names placement will optimize feature searches. Point features are relatively easy to associate with their names, since point symbols are often recognizable by software and lie within a reasonable distance from their labels. Some linear features can be automatically associated to their names if their symbols have identifiable characteristics (e.g., distinctive color, shape, or screen). It may also be possible for the system to infer which type of feature to look for, based on type style, type color, or a generic component of the name.

The map image and ASCII names are displayed to the analyst with feature associations highlighted for correction. Then, the analyst processes unassociated names.

4.8 Position Capture

The system automatically digitizes the locations of point features associated to words. Software computes each point symbol's mean center. For point features whose symbols' centers of gravity do not correspond to precise position, feature position relative to symbol position must be stored in a symbol table.

Most line and area feature processing and quality control of point feature processing is left to the analyst and his cursor. For quality control, the software echos the analyst's input by placing a small symbol at digitized points on the map background. Bounded areal features (incorporated towns, lakes, political entities) are digitized at their centroid, generalized boundary points, and points where they leave the map sheet. If a large portion of an areal feature is off the map, its digitized centroid should be adjusted accordingly. Linear features are digitized at both ends or at map sheet entry or exit points.

Some names refer to locales or features with indefinite boundaries (e.g., bays, forests, deserts). In such cases, the analyst must use his own judgment. Locale names can be tied to groups of buildings, a crossroad, or the location at the name's center. Other areal features are centered on the name, if boundaries are indeterminate.

Several upgrade possibilities for position capture software exist. When areal features are symbolized by fills or screens, it may be possible to create a crude boundary based on the extent of the fill or to construct a minimum bounding rectangle to describe the feature's areal extent. Some areal features, especially natural features (swamps, deserts, bogs), may consist of more than one region or polygon, in which case more than one centroid or boundary file would be associated with the same name. Island chains must be represented as a polygon that includes all the islands without attention to individual island position. Line-following software could be implemented to determine the limits of linear features and of areal features bounded by line symbols. None of the options discussed in this paragraph, however, are cost-effective at this development stage.

4.9 Attribute Capture

The initial system relies on input from the analyst to capture feature attributes. All possible attributes are captured for all named features. Usually, attribute information is acquired from nonmap sources.

A number of upgrade possibilities exist; most, however, are effective only on DMA or other standard product series. Scaled point symbol size, particularly populated place symbols, indicate categories of feature size. Type fonts can indicate feature classes, but these are not consistently reliable. The military or strategic significance of places is sometimes indicated by point symbols adjacent to the placename. If such symbols are used consistently, it could be useful to recognize them automatically. Road classes are represented by unique line symbols. However, few roads are named. Size and importance of areal features is indicated by type font and by the symbol's areal extent. These and other upgrade possibilities should be evaluated for cost-effectiveness before being developed. In many cases, neither the quality nor the quantity of information gained justifies the development expense.

4.10 Quality Control

Each of the functions described above must include automated checks to catch extreme or unlikely data values. Rules must be defined to filter names, positions, feature classes, and attributes that are clearly in error. Error conditions are communicated to the analyst, along with all relevant information.

After a map is processed, a proof plot and a data listing are generated from the captured names data to verify the digitized data against the source(s). The proof plot must include registration marks, names, positions of the named features, and an indication of feature types. The proof plot must match the source map's projection and scale. The data listing must show the entire data record of each captured name. It should include flags to indicate problem records and missing fields.

The analyst must be able to display and edit any field of any name on the data listing, and graphically edit the proof plot at the workstation. This stage of quality control ascertains that the captured names data truthfully reflect the source data and corrects registration, scale, and projection errors. Toponymic evaluation of the names data follows.

4.11 Output Processing

A Data Base Update File is created from the captured data and is scheduled for entry to the data base. File generation is reported to the data base manager.

4.12 File Management

File management must be able to create, manipulate, and use as input AADES workfiles, lookup tables, accounting files, and output files. Also necessary is the ability to load and link programs. Relocation functions for manipulating programs according to storage allocation and load functions for loading these programs into storage are needed. The ability to link independently translated programs must also exist. Such programs include operating system modules, a general user library function, user object programs, and others.

4.13 Job Management

Job management controls job initiation and termination and within-job processes. Job management must

- establish and terminate accounting procedures for a job,
- associate I/O device addresses with symbolic addresses,
- invoke systems programs to perform requested tasks,
- control and allocate system resources efficiently.

5.0 DATA SETS

5.1 Lookup Tables

The system accesses lookup tables during processing to obtain product-specific or area-specific parameters. Which lookup tables are accessed is based on attributes of the source map (input at the preprocessing stage). Symbol tables and type style tables are selected according to the type of product being processed. The product's area and its language dictate use of diacritics dictionaries. Stored font libraries must be accessed by template-matching OCR subroutines.

AADES stores a Product Specification Table for each DMA standard product. Once the analyst inputs the product type during preprocessing, the system automatically selects projection conversion software, adjusts for scale, and chooses the appropriate feature symbol and type style assignment tables.

5.2 Workfiles

5.2.1 Map Image File

The imaged map is stored in a workfile in bit planes relating to color separations. The raster data file is manipulated by interactive and batch subroutines. It may be segmented, mosaicked, enhanced, or displayed, in addition to providing input to names capture subroutines.

5.2.2 In-Process Workfiles

Data structuring will dictate the number and nature of such files. At each stage of processing, successively more complex information must be stored. All data must be readily accessible to the system and the analyst. All such files must be savable, so interactive processing can resume after interruption.

5.2.3 Names Data File

As names data are captured they are organized into names data records that contain all captured names and attributes.

5.3 Data Base Update File

Output processing generates a Data Base Update File, which employs standard data transfer format (Fig. 5-1). Data Base Update Files are routed to toponymists for supervised entry into the data base.

Table 5-1. Standard Data Transfer Record.

<u>Entity Name</u>	<u>Size (Bytes)</u>
Data Source Name	10 (1)
Number of Characters in Geoname	1
Number of Characters in Non-Anglicized Name	1
Number of Characters in Alias	1
Number of Characters in Province Name	1
Number of Characters in Country Name	1
Names (geoname, non-Anglicized name, alias, province name, country name)	140 (2)
Type of Romanization	1
Date of Data Source	3 (3)
Date of Data Capture	3
Date of Last Update	3
Position	6 (4)
Positional Accuracy	2
Feature Designator	6
Attribute	6
Administrative Code	1
Area Code	1
UTM grid	8
Selected Map Sheet	7
Approved or not Approved	1
Bounding Rectangle	13 (5)
Pointer to File Containing Feature Coordinates	8
Unused	32
	<hr/> 256

(1) The GNDB maintains a dictionary of legal data sources.

(2) If more than 140 characters are required, the next record is an overflow record. All names are stored in this field to substitute one large field with overflow allowances for potentially five large fields with possible overflows.

(3) Dates are numeric strings: ddmmyy.

(4) Position as currently planned is a point (the location of a point feature, the mouth of a river, or the centroid of an area feature) given as two signed numeric strings: +/- dddmmss and +/- dddmmss. Negative indicates latitude South or longitude East, positive indicates latitude North or longitude West.

(5) The bounding rectangle is high and low latitudes and longitudes, with an additional byte indicating if the bounding rectangle is incomplete due to the feature leaving the map.

APPENDIX A

PREVIOUS WORK IN NAMES DATA CAPTURE

This section describes a number of available names data capture methods in current or past use, or under development.

A.1 National Geographic Society

The National Geographic Society generated a names data base to provide correctly spelled names for user-specified locations or other parameters. Later system upgrades will perform typesetting and map annotation.

To enter names, one of five different coding sheets is completed by a researcher. Coordinates in degrees and minutes are manually calculated. The coded information is keystroked at a computer terminal (IBM 3278). Names with diacritics are entered using special terminal keyboard overlays.

DMA's prototype Names Input Station (NIS) already exceeds the technology used by the National Geographic Society. The NIS's digitizing table is a far more efficient means of capturing geographic locations.

A.2 U.S. Geological Survey

To maintain an authoritative on-line file of geographic place names, the USGS Geographic Names Branch has captured all but the road and highway names appearing on its maps. Future plans are to capture names from other sources and to capture road and highway names. An interface to an automated map names processing system is also planned.

The USGS's data capture was contracted to a small business minority contractor. Preparation of maps prior to digitizing, however, was performed by in-house personnel. This preprocessing step was arduous and time-consuming, but its costs are not reflected in project cost summaries (in-house labor was not included). The contractor used a digitizing table and a keyboard to enter names from maps marked by USGS personnel. Payment was on a per-name basis.

Time figures for preparing Arkansas placenames for capture were provided by the USGS (Table A-1). Forty hours were devoted by USGS personnel to assembling, organizing, and marking USGS Arkansas maps. Names to be digitized and their feature locations were marked on the maps. The contractor subsequently spent 20 hours sorting the maps according to his processing scheme. Once sorted, capture involved selecting a name, locating its feature, digitizing the point, and typing the name and its attributes (in this case, feature type, area code, and map number). Following capture, the digitized data were verified.

DMA's NIS provides the capability to digitize names in this manner. However, it is evident from USGS cost estimates and from DMA experience that many years will pass before the data base is completely loaded without more personnel or greater input speed.

Table A-1. USGS data capture rate. The 27,684 Arkansas names data records (whose capture time is shown below) required approximately 3.54 minutes each to capture.

	<u>Hours</u>
Preparation	40
Data entry (1)	
Sort (2)	20
Annotate	300
Verify	108
Key	424
Edit	110
Correct keying mistakes	24
Processing (background)	20
Administration time	88
Additional corrections	8
Total data entry	1102
Quality control (3)	155
Correction	250
Editing	85
Administrative	<u>3</u>
	1635

(1) Data captured are name, feature class, FIPS County Code, positional coordinates, Map Code, and elevation.

(2) Maps, which arrived in alphabetical order by title, were sorted into 10 X 10 cells for more efficient data entry.

(3) Plots were made of captured names and compared to map sheets.

A.3 Chicago Aerial Survey

The Chicago Aerial Survey uses an interactive graphics workstation to digitize names data. Names and feature attributes are keystroked prior to position capture. Digitized name and feature data are moved to a disk that is accessed by an Intergraph workstation. Using a map fastened to the workstation table, coordinates are digitized and linked to their respective name information.

While sophisticated and expensive hardware is used in this method, input speed is not appreciably improved. Analyst comfort, however, is improved by the ergonomically designed workstation. This method is likely to increase keystroking speed, since dedicated keystrokers enter data without interruption. Increased keystroking speed, however, is offset by the time required to search the map for each name's feature position.

A.4 Central Intelligence Agency

Using a map fastened to a workstation table (Intergraph), the system-specific interactive utilities are invoked to digitize locations and keystroke names and feature attributes. Following entry, data are transferred to a mainframe computer, plotted, and listed. These materials are reviewed and corrections are performed interactively at the workstation.

If low technical risk and immediate procurement are important AADES requirements, this system with added diacritics capabilities is recommended. However, data entry speed would be only marginally better than the prototype Names Input Station, with limited upgradability.

A.5 Scitex and Intergraph

Both Scitex and Intergraph claim softcopy-based map reading capabilities. A map separation is scanned, text is indicated by an analyst, and the text image, position, and orientation are recorded by the system. Text images are compared to a type font library. If recognized, the ASCII code is stored in the data base with position and orientation. If unrecognized, the ASCII code is supplied interactively. All attribute information is entered interactively.

These systems fit the AADES functional description in Section 4. No project personnel, however, have witnessed their operation, and some AADES requirements are not addressed. In particular, AADES must read diacritics and special symbols, as well as provide keyboard entry of such marks. Also, AADES must process one-of-a-kind maps whose type styles are not in a system lookup table. Inferential character recognition algorithms are required by DMA to supplement template matching or OCR for non-series maps will not be successful.

APPENDIX B

REFERENCES

"Apache TM Functional Specifications," Altek Corp., Silver Spring, Maryland, August 1983.

"Auerback on Optical Character Recognition," Auerback Computer Technology Reports, June 1982.

"Automation of the National Toponymic Data Base," by J. S. Thompson of Energy, Mines, and Resources, Canada, October 10, 1979.

Brown, R. M. and C. F. Cheng, *"Optical Character Recognition for Automated Cartography: The Advanced Development Handprinted Symbol Recognition System,"* Naval Ocean Research and Development Activity, NSTL, Mississippi, NORDA Technical Note 187, March 1983.

Brown, Robert, *"Preprocessing for Symbol Recognition,"* Naval Ocean Research and Development Activity, NSTL, Mississippi, NORDA Technical Note 210, February 1985.

"Feature Categories of the Geographic Names Information System (GNIS)," U. S. Geological Survey, December 1983.

Gronmeyer, Larry K., B. W. Ruffin, M. A. Lybanon, S. E. Pierce, and P. L. Neely, *"An Overview of Optical Character Recognition (OCR) Technology Techniques,"* Naval Ocean Research and Development Activity, NSTL, Mississippi, NORDA Technical Note 217, June 1978.

Hansen, Erik, *"Place Name Register,"* Danish Society of Cartography, November 1979.

Horner, W. R. and S. P. Schumacher, *"Implications of Symbol Usage on U. S. Army Maps for an Automated Cartographic System,"* and *"Technical Appendices to ...,"* Engineering Topographic Labs, Ft. Belvoir, Virginia.

Jablinske, R., D. Strife, K. Gaar, and J. Moore, *"Names Type File System,"* DoD Electromagnetic Compatibility Analysis Center, Annapolis, Maryland, April 1983.

Nagy, George, *"Criteria for Selecting Automatic Digitizers (Optical Scanners),"* University of Nebraska, Lincoln, Unpublished report, 1982.

Orth, Donald J. and Roger L. Payne, *"The National Geographic Names Data Base: Phase II Instructions,"* U.S. Geological Survey Open File Report 84-036, Reston, Virginia, 1984.

Poiker, Thomas K., *"An Intelligent Cursor," Computer Graphics World*, pp. 91-93, October 1983.

"Report of Findings of Automated Name Placement Research: Phase I," Unit Basoglu, CACI, Inc., Federal for USGS, September 1982.

"Standing Operating Procedures for Servicing, Maintaining, and Disseminating Names, Boundaries, Populated Place Classifications, and Miscellaneous Items on Topographic and Hydrographic Products," DMA, December 1980.

Walker, Charles, Robert Brown, and Walter Osterman, *"Raster Scan Character Recognition System,"* Naval Ocean Research and Development Activity, NSTL, Mississippi, NORDA Technical Note 188, March 1983.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

AD-A162 458

REPORT DOCUMENTATION PAGE																
1a. REPORT SECURITY CLASSIFICATION Unclassified		1b. RESTRICTIVE MARKINGS None														
2a. SECURITY CLASSIFICATION AUTHORITY		3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.														
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE																
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NORDA Report 98		5. MONITORING ORGANIZATION REPORT NUMBER(S) NORDA Report 98														
6. NAME OF PERFORMING ORGANIZATION Naval Ocean Research and Development Activity		7a. NAME OF MONITORING ORGANIZATION Naval Ocean Research and Development Activity														
6c. ADDRESS (City, State, and ZIP Code) Ocean Science Directorate NSTL, Mississippi 39529-5004		7b. ADDRESS (City, State, and ZIP Code) Ocean Science Directorate NSTL, Mississippi 39529-5004														
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Defense Mapping Agency	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER														
8c. ADDRESS (City, State, and ZIP Code) HQ/STT Washington DC 20305		10. SOURCE OF FUNDING NOS. <table border="1"><tr><td>PROGRAM ELEMENT NO 64710B</td><td>PROJECT NO.</td><td>TASK NO.</td><td>WORK UNIT NO.</td></tr></table>			PROGRAM ELEMENT NO 64710B	PROJECT NO.	TASK NO.	WORK UNIT NO.								
PROGRAM ELEMENT NO 64710B	PROJECT NO.	TASK NO.	WORK UNIT NO.													
11. TITLE (Include Security Classification) The Geonames Processing System Functional Design Specification, Volume 1: Automated Alphanumeric Data Entry System																
12. PERSONAL AUTHOR(S) Gail Langran, Anne Downs, Barry Glick, and Warren Schmidt																
13a. TYPE OF REPORT Final	13b. TIME COVERED From _____ To _____	14. DATE OF REPORT (Yr., Mo., Day) March 1985	15. PAGE COUNT 26													
16. SUPPLEMENTARY NOTATION																
17. COSATI CODES <table border="1"><tr><th>FIELD</th><th>GROUP</th><th>SUB GR</th></tr><tr><td> </td><td> </td><td> </td></tr><tr><td> </td><td> </td><td> </td></tr><tr><td> </td><td> </td><td> </td></tr></table>		FIELD	GROUP	SUB GR										18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) Maps, computers, software systems		
FIELD	GROUP	SUB GR														
19. ABSTRACT (Continue on reverse if necessary and identify by block number) This report describes the Geonames Processing System attributes and serves as a basis for understanding between the user and the developer. The subsystems referred to are: Advanced Symbol Processing, Advanced Type Placement, Geographic Names Data Base, and Automated Alphanumeric Data Entry System. (P 1-1)																
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT <input checked="" type="checkbox"/> DTIC USERS <input type="checkbox"/>		21. ABSTRACT SECURITY CLASSIFICATION Unclassified														
22a. NAME OF RESPONSIBLE INDIVIDUAL Gail Langran		22b. TELEPHONE NUMBER (Include Area Code) (601) 688-4449	22c. OFFICE SYMBOL Code 351													

DD FORM 1473, 83 APR

EDITION OF 1 JAN 73 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

END

FILMED

2-86

DTIC